A Scalable, Commodity Data Center Architecture

1

Overview

- Structure and Properties of a Data Center
- Desired properties in a DC Architecture
- Fat tree based solution
- Monsoon: layer 2 flat routing

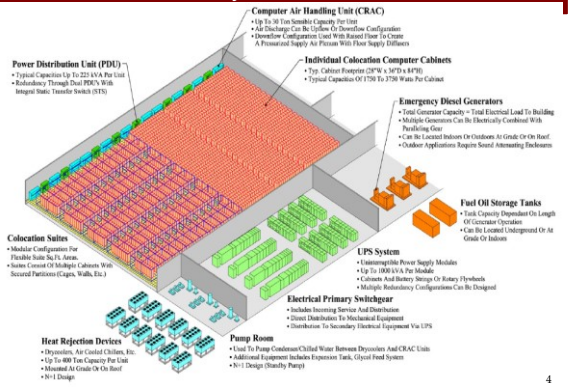
2

Anatomy of a Datacenter

- Servers (Physical machines)
- Storage
- Network devices (switch, router, cables)
- Power distribution systems
- Cooling systems
- …

3

Anatomy of a Datacenter



4

Electrical Power-1

- Condition circuits
  - Little power fluctuations
- UPS
  - On Line
    - o Power is conditioned by the unit
    - o Equipment draws power from UPS all the time
  - Off Line
    - o Equipment draws power from the UPS only when external source of electricity has been lost
- UPS have short power cycle
  - Back up with Generators
- Automatic Transfer Switch – ATS
- Compute the Electrical Load
  - All Components, UPS charging

5

Electrical Power-2

- Overhead power distribution
- Racks with Power Distribution Units (PDU)
  - Keep it all within the rack
- Multiple power sources (circuits)

6

A DC-wide System

- Has software systems consisting of:
  - Distributed system, logical clocks, coordination and locks, remote procedural call…etc
  - Distributed file system
  - (We do not go deeper into above components)
  - Parallel computation: MapReduce, Hadoop
- Virtualized Infrastructure:
  - Computing: Virtual Machine / Hypervisor
  - Storage: Virtualized / distributed storage
  - Network: Network virtualization

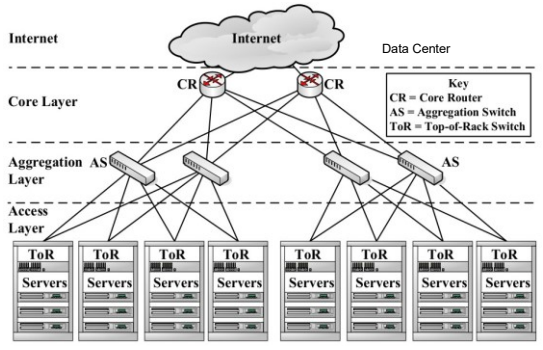
7

Conventional Topology

- Three layers:
  - Access layer with Top of the Rack (ToR) switches
  - Aggregation layer
  - Core layer

8

Common data center topology



9

Virtualized Data Center

Data center with some or all the hardware virtualized

- Servers (Physical machines)
- Storage
- Network devices (switch, router)
- Power distribution systems
- Cooling systems

10

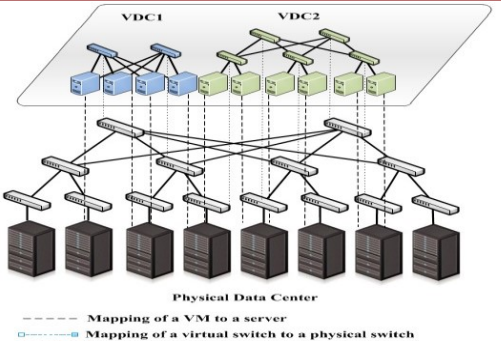
Virtual Data Center

Collection of virtual resources, e.g.

- Virtual machine
- Virtual switches
- Virtual links

11

Virtual Data Center



12

Problem With common DC topology

- Single point of failure
- Over subscript of links higher up in the topology
  - Trade off between cost and provisioning

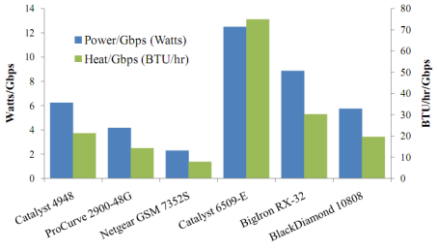
13

Properties of solutions

- Backwards compatible with existing infrastructure
  - No changes in application
  - Support of layer 2 (Ethernet)
- Cost effective
  - Low power consumption & heat emission
  - Cheap infrastructure
- Allows host communication at line speed

14

Cost of maintaining switches



15

Need for Layer 2 In DC

- Certain monitoring apps require server with same role to be on the same vlan
- Using same ip on dual homed servers
- Allowing growth of server farms.

16

Review of Layer 2 & Layer 3

- Layer 2
  - One spanning tree for entire network
    - Prevents looping
    - Ignores alternate paths
- Layer 3
  - Shortest path routing between source and destination
  - Best-effort delivery

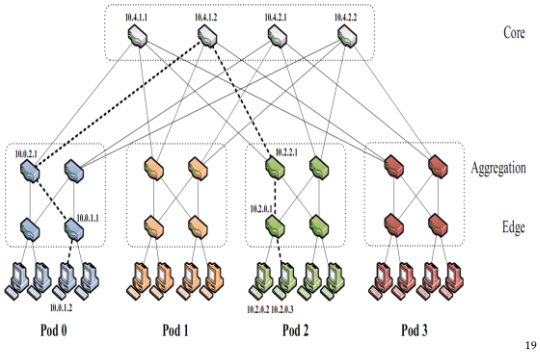
17

FAT Tree based Solution

- Connect end-host together using a fat tree topology
  - Infrastructure consist of cheap devices
    - Each port supports same speed as endhost
  - All devices can transmit at line speed if packets are distributed along existing paths
  - A k-port fat tree can support k3/4 hosts

18

Fat-Tree Topology



19

Problems with a vanilla Fat-tree

- Layer 3 will only use one of the existing equal cost paths
- Packet re-ordering occurs if layer 3 blindly takes advantage of path diversity

20

FAT-tree Modified

- Enforce special addressing scheme in DC
  - Allows host attached to same switch to route only through switch
  - Allows inter-pod traffic to stay within pod
  - unused.PodNumber.switchnumber.Endhost
- Use two level look-ups to distribute traffic and maintain packet ordering.

Prefix	Output port
10.2.0.0/24	0
10.2.1.0/24	1
0.0.0.0/0	

Suffix	Output port
0.0.0.2/8	2
0.0.0.3/8	3

21

2 Level look-ups

- First level is prefix lookup
  - Used to route down the topology to endhost
- Second level is a suffix lookup
  - Used to route up towards core
  - Diffuses and spreads out traffic
  - Maintains packet ordering by using the same ports for the same endhost

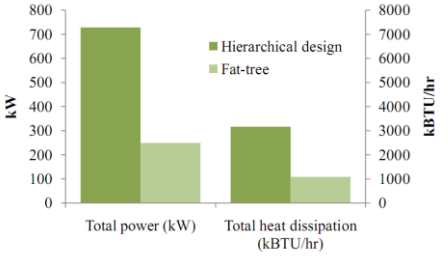
22

Diffusion Optimizations

- Flow classification
  - Eliminates local congestion
  - Assign to traffic to ports on a per-flow basis instead of a per-host basis
- Flow scheduling
  - Eliminates global congestion
  - Prevent long lived flows from sharing the same links
  - Assign long lived flows to different links

23

Results: Heat & Power Consumption



24

Draw Backs

- No inherent support for VLan traffic
- Data center is fixed in size
- Ignored connectivity to the internet
- Waste of address space
  - Requires NAT at border

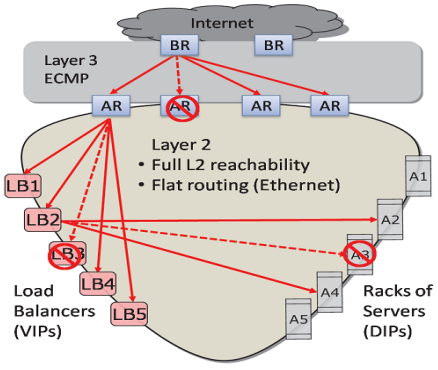
25

Monsoon approach

- Layer 2 based using future commodity switches
- Hierarchy has 2:
  - access switches (top of rack)
  - load balancing switches
- Eliminate spanning tree
  - Flat routing
  - Allows network to take advantage of path diversity
- Prevent MAC address learning
  - 4D architecture to distribute data plane information
  - TOR: Only need to learn address for the intermediate switches
  - Core: learn for TOR switches
- Support efficient grouping of hosts (VLAN replacement)

26

Monsoon



27

Monsoon Components

- Top-of-Rack switch:
  - Aggregate traffic from 20 end host in a rack
  - Performs ip to mac translation
- Intermediate Switch
  - Disperses traffic
  - Balances traffic among switches
  - Used for valiant load balancing
- Decision Element
  - Places routes in switches
  - Maintain a directory services of IP to MAC
- Endhost
  - Performs ip to mac lookup

28

How routing works

- End-host checks flow cache for MAC of flow
  - If not found ask monsoon agent to resolve
  - Agent returns list of MACs for server and MACs for intermediate routers
- Send traffic to Top of Router
  - Traffic is triple encapsulated
- Traffic is sent to intermediate destination
- Traffic is sent to Top of rack switch of destination

29